Learning User Attributes via Mobile Social Multimedia Analytics

LIQIANG NIE, Shandong University, China
LUMING ZHANG, MENG WANG, and RICHANG HONG, Hefei University of Technology, China
ALEKSANDR FARSEEV and TAT-SENG CHUA, National University of Singapore, Singapore

Learning user attributes from mobile social media is a fundamental basis for many applications, such as personalized and targeting services. A large and growing body of literature has investigated the user attributes learning problem. However, far too little attention has been paid to jointly consider the dual heterogeneities of user attributes learning by harvesting multiple social media sources. In particular, user attributes are complementarily and comprehensively characterized by multiple social media sources, including footprints from Foursqare, daily updates from Twitter, professional careers from Linkedin, and photo posts from Instagram. On the other hand, attributes are inter-correlated in a complex way rather than independent to each other, and highly related attributes may share similar feature sets. Towards this end, we proposed a unified model to jointly regularize the source consistency and graph-constrained relatedness among tasks. As a byproduct, it is able to learn the attribute-specific and attribute-sharing features via graph-guided fused lasso penalty. Besides, we have theoretically demonstrated its optimization. Extensive evaluations on a real-world dataset thoroughly demonstrated the effectiveness of our proposed model.

Categories and Subject Descriptors: H.3.3 [**Information Systems**]: Information Search and Retrieval

General Terms: Algorithms, Experimentation, Performance

Additional Key Words and Phrases: Learning user attributes, mobile social multimedia analytic, occupation inference

ACM Reference Format:
Liqiang Nie, Luming Zhang, Meng Wang, Richang Hong, Aleksandr Farseev, and Tat-Seng Chua. 2016. Learning user attributes via mobile social multimedia analytics. ACM Trans. Intell. Syst. Technol. 8, 3, Article 36 (April 2017), 19 pages.
DOI: http://dx.doi.org/10.1145/2963105

1. INTRODUCTION

Social media and mobile phones promote mutually. In particular, the development of social networking technologies has changed the role of mobile phones, which, apart from pure communication devices, are also powerful devices for generating and consuming multimedia data [Ji et al. 2014]. The growing ubiquity of mobile phones has

This research was supported by "Qilu Scholar" grant.
Authors' adresses: L. Zhang, M. Wang, and R. Hong are with the Department of Computer Sciences and Information Engineer, Hefei Unviersity of Technology, 193 Tunxi Road, Hefei, Anhui, P.R. China (230009); emails: {zglumg, eric.mengwang, hongrc.hfut}@gmail.com; L. Zhang is also with National University of Singapore (Suzhou) Research Institute 377 Lin Quan Street, Suzhou Industrial Park, Jiang Su, China; L. Nie is with the Department of Computer Science and Technology, Shandong University, No.1500, Shunhua Road, Jinan, Shandong Province, P.R. China (250101); email: nieliqiang@gmail.com; He is the corresponding author (email: nieliqiang@gmail.com); A. Farseev and T.-S. Chua are with the School of Computing, National University of Singapore, AS6, $05-25, Computing 1, 13 Computing Drive Singapore (17417); emails: farseev@gmail.com, chuats@comp.nus.edu.sg.

DOI: http://dx.doi.org/10.1145/2963105

in turn led to a bright new stage of social multimedia services. In light of this, users can conveniently capture images and videos on the mobile end, associate them with social and contextual metadata such as tips and GPS tags, and disseminate them in social platforms. Hence, social media and mobile users constitute an ideal platform for analyzing users' habits and behaviors, which also provide us the most influential sources in shaping user attributes, such as age, gender, occupation, and social status.

User attributes are crucial prerequisites for many interesting applications. We list a few examples. (1) Studying social roles and statuses is very helpful to gain insights into the whole society as well as manage social resources at the individual level; (2) age and gender inference substantially contribute to analyzing users' profiles and conducting demographic statistics; and (3) interest and occupation prediction benefit customized marketing and personalized recommendation, as well as social circle detection and activity recognition. Noticeably, a significant body of Internet users might be reluctant to expose their attributes to the public. As an alternative way, predicting user attributes from mobile social media is of great interest to both industry and academia.

Despite its value and significance, learning user attributes from mobile social media remains in its infancy due to the following challenges: (1) People may be involved in multiple mobile social media platforms for various purposes simultaneously.[1] For example, people share their footprints with their friends using Foursquare[2]; meanwhile, they may also share the latest news using Twitter[3] and photos using Instagram.[4] Different aspects of users are disclosed on different social networks due to their different emphases. However, these heterogeneous multimedia sources are complementary to each other and essentially characterize the same user from different perspectives. Effectively unifying and uncovering the information embedded in the heterogeneous social media sources remains a largely unaddressed research problem. (2) User attributes are typically correlated in a non-uniform way. Take a user's occupation prediction as an example. Given a set of occupations, $C = \{nurse; dentist; scientist; professor, cook\}$, the relatedness between nurse and dentist may be stronger than that between nurse and cook, since nurse and dentist work in similar environments, and they hence may discuss similar topics in their social forums and post images with similar context. (3) Features describing users may suffer from the curse of dimensionality, but in fact not all features are discriminant. In addition, a subset of highly related attributes may share a common set of features, whereas weakly related attributes are less likely to be affected by the same features. In summary, learning attribute-sharing features and attribute-specific features effectively is significant to user attributes learning. These together pose crucial challenges for us.

To address the aforementioned challenges, we propose a unified model, the so-called graph-constrained multi-source multi-task learning model (*gMM*), to infer user attributes. As a research entry point, we specifically consider occupation prediction in this work. We assume that each user is involved in multiple social media platforms.[5] Our model treats each occupation as a task. It simultaneously co-regulates source consistency and task relatedness. In particular, for each given user, we first crawl his/her historical multimedia posts from multiple mobile platforms, including tweets from Twitter, check-in records from Foursquare, images from Instagram, as well as occupations from Linkedin (ground truth). The occupations revealed by different social

[1]It is reported that 52% of online adults in 2014 use multiple social media sites: http://www.pewinternet.org/2015/01/09/social-media-update-2014/.

[2]https://foursquare.com/.

[3]https://twitter.com/.

[4]https://instagram.com/.

[5]Even though some users do not have multiple social platforms, we can utilize matrix factorization techniques to complete the missing sources.

networks for the same users should be similar, and the inconsistency among the prediction results from individual sources should be penalized. On the other hand, the relatedness among occupations is encoded with the graph-guided fused lasso, where each task is represented by a node in a graph and the similarities between the tasks are captured via an adjacency matrix. The graph structure is constructed based on external and internal knowledge. As compared to the most previous methods regrading all task pairs as equally related and cannot encode the hidden complicated relationships among tasks, our method uses the graph-guided fused lasso penalty to encourage models for task pairs to be similar whenever they are connected in the graph network. In addition, the proposed *gMM* model is capable of identifying discriminant task-specific and task-sharing features. As an added benefit, it is advantageous to learn tasks by leveraging cues from the other related tasks, especially when the data are scare. It can thus alleviates the problem of insufficient training samples. Also, our model can be generalizable to other attributes, such as interests, age groups, and social roles.

The contributions of this work can be summarized as follows:

—We proposed a multi-source learning model with the graph-guided fused lasso penalty to infer user occupations, which jointly regularizes source consistency and task relatedness.
—We relaxed the non-smooth objective function to a smooth one, theoretically demonstrated its solution and practically analyzed its computational complexity;
—We constructed a representative dataset by crawling multiple mobile social media platforms, extracted descriptive features from multiple sources, and validated our proposed model on that.

The remainder of the article is structured as follows. Section 2 reviews the related work. In Section 3, we introduce the procedure of data construction. Sections 4 and 5 detail our proposed model and its optimization, respectively. Experimental settings and results are reported in Section 6, followed by conclusion and future works in Section 7.

2. RELATED WORK

2.1. User Attributes Learning from Social Media

Mobile social media platforms provide venues for diverse people to record and share their behaviors. Their online behaviors are representative of many aspects of their attributes. Hence predicting user attributes from social media is feasible and draws more and more attention. Recent works focus on the inference of age [Rosenthal and McKeown 2011], gender [Dong et al. 2014], race [Nguyen and Lim 2014], occupation [9], personality [Gou et al. 2014], and political alignment [Conover et al. 2011]. For example, the authors in Burger et al. [2011] constructed a large, multilingual dataset labeled with gender and studied several statistical models for determining the gender of uncharacterized twitter users. The work in Pennacchiotti and Popescu [2011] attempts to automatically infer the political orientation and ethnicity of given users by leveraging observable information such as user behaviors, network structures, and the linguistic contents of the users' Twitter feeds. The work introduced in Rao et al. [2010] discovers four latent user attributes, including gender, age, regional origin, and political orientation, by use of a stacked Support Vector Machine– (SVM) based classification algorithm over a rich set of original features extracted solely from informal content of Twitter. Furthermore, the authors in Zamal et al. [2012] extended the existing work on attribute inference by leveraging the principle of homophily. They evaluated the inference accuracy gained by augmenting the user features with features derived from the Twitter profiles and posts of friends. They considered three attributes that have varying degrees of assortativity: gender, age, and political affiliation. Recently, the authors in He and Lawrence [2011] noticed that existing approaches for user attributes

learning on social media are generally in supervised settings and they formulated a weakly supervised paradigm to extract user profiles from Twitter.

Even though considerable success has been achieved previously, most of them learn user attributes from single source and few explore the relatedness among the attributes. In fact, we are living in an era of multiple social networks. Users are using more and more social networks simultaneously. It is hence desirable to integrate multiple social media platforms to comprehensively characterize the same users. On the other hand, the relatedness among attributes can be learned from external knowledge or can be manually defined based on prior knowledge. As an improved work, our model incorporates these two factors.

2.2. Multi-View Multi-Task Learning

The problem of user attributes learning from multiple social media platforms exhibits dual heterogeneities: Every task in the problem has features from multiple sources, and multiple tasks are related to each other in a complex way. Hence, multi-view multi-task learning would be a natural and an optimal model for user attributes learning. In retrospect, there are only a few studies in the literature on multi-task problem with multi-source data, and a few of them have been applied to user attributes learning. He and Lawrence [2011] proposed an iterative framework for multi-view multi-task learning (IteM^2) with its applications to text classification. IteM^2 projects task pairs to a new Reproducing Kernel Hilbert Space based on the common views shared by them. However, it is specifically designed to handle non-negative feature values. Even worse, as a transductive model, it fails to generate predictive models on independent and unknown samples. To address the intrinsic limitations of transductive models, an inductive multi-view multi-task learning model regMVMT was introduced in Zhang and Huan [2012]. regMVMT uses co-regularization to obtain functions consistent with each other on the unlabeled samples from different views. Across different tasks, additional regularization functions are utilized to ensure that the learned functions are similar. However, simply assuming all tasks are similar without prior knowledge might be inappropriate. As a generalized model of regMVMT, an inductive convex shared structure learning algorithm for the multi-view multi-task problem (CSL-MTMV) was developed in Jin et al. [2013]. Beyond regMVMT, CSL-MTMV considers the shared predictive structure among multiple tasks.

Our proposed model differs from the above methods from three perspectives: (1) The existing methods maximize the agreement between views using unlabeled data, while *gMM* does not assume that there are abundant unlabeled data. (2) IteM^2, regMVMT, and CSL-MTMV are all binary classification models, which require nontrivial extensions in order to handle multi-class problems, especially when the number of classes is large. Our model can jointly learn multiple classes. (3) Our model is able to learn the task-sharing features and task-specific features using lasso and graph-regularized fused lasso techniques.

3. DATA CONSTRUCTION

3.1. Data Crawling Strategy

In this section, we discuss multiple sources construction by crawling data of the same users from their multiple mobile social media platforms. The challenge is how to align the same users across different social media platforms. Towards this end, we leveraged the emerging social services About.me.[6] The site offers registered users a convenient platform to link multiple online identities, relevant external sites, and popular social

[6]https://about.me/.

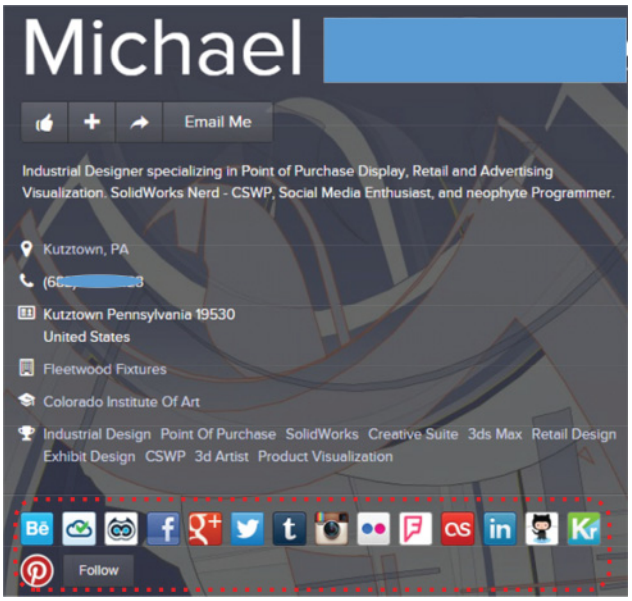

Fig. 1. Illustration of social account alignment via About.me. The icons in the red dotted rectangle are links to other social accounts of the same person. To avoid the leakage of private information, we deliberately partially cover the name and contact information.

networking websites such as Foursquare, Instagram, LinkedIn, Twitter, and Flickr.[7] It is characterized by its one-page user profiles, each with a large and artistic background image and abbreviated biography. Figure 1 elaborates the accounts alignment strategy via About.me.

We utilized the following strategy to construct our multi-source data:

(1) We searched About.me with some manually selected occupation concepts. These occupations were selected from this alphabetical list.[8]
(2) We abandoned those users who fail to list the following three social accounts in their About.me profile: Twitter, Linkedin, and Foursquare. For each of the rest users, we have his/her three URLs corresponding to the three social accounts. According to this criteria, we collected a set of 3,180 users.
(3) For each of these 3,180 users, we collected all their historical posts from Linkedin, Twitter, and Foursquare, respectively.
(4) We also collected images from Instagram for these 3,180 users using the following procedure: If users provide their Instagram URL on About.me, then we can directly crawl their images[9]; otherwise, based on the Foursquare check-ins, we collected photos related to the visiting venues from Instagram. The venue is matched according to the identity between Foursquare and Instagram.[10] Users some times share their Instagram photos in their Twitter account with the snippet "instagram.com," and we thus collected them to enrich our image collection.

Table I shows the first-order statistics of our collected data from multiple sources. It is observable that each user on average posted more than 2,300 tweets. This indicates

[7]https://www.flickr.com/.
[8]http://www.occupationsguide.cz/en/abecedni/abecedni.htm.
[9]We observed 40.9% of these 3,180 users also explicitly post their Instagram access in their About.me profile.
[10]http://instagram.com/developer/endpoints/locations/.

Table I. Statistics of Our Collected Data from Multiple Sources, Including Tweets from Twitter, Check-ins, and Tips from Foursquare, Profiles from Linkedin, and Photos from Instagram

Data Sources	Users	Tweets from Twitter	Check-ins from Foursquare	Tips from Foursquare	Profiles from Linkedin	Images from Instagram
Statistics	3,180	7,365,373	28,504	22,079	3,180	106,684

Table II. This Table Lists 20 Representative Occupations in our Dataset. In Fact, We Studied 80 Occupations. to Save Space, We Do Not List Them All

ID	Occupations	ID	Occupations	ID	Occupations	ID	Occupations
1	Sales/Markets	6	Researcher	11	Writer	16	Photograher
2	Software Engineer	7	CEO	12	Waiter	17	Product Manager
3	Founder/Cofunder	8	Project Manager	13	Artist	18	Teacher
4	Consultant	9	Student	14	Blogger	19	Editor
5	Web Designer	10	Agent	15	CTO	20	Investor

that the users with multiple social media accounts at the same time are usually very active. These four sources characterize users from different views. In particular, Linkedin, a professional networking platform, serves as the source of ground truth, where users typically post their professional and career information; Twitter is an online social networking service that enables users to send and read short 140-character messages called "tweets." Some of these tweets may signal occupation-related information. For instance, property agents may talk about rentals, stamp duty, and commission rates; Foursquare records users' travel histories, which can implicitly leak out the behaviors or patterns of one's profession. Take professors and bankers as examples. Check-ins of professors primarily distribute in universities and conference venues; while the check-ins for bankers usually occur in financial institutions and central business districts. Images in Instagram can visually and intuitively characterize the users' working places.

3.2. Ground-Truth Construction

We leveraged the structural information of users' Linkedin profiles to establish the ground truth, which greatly saves us from the labour-intensive labeling process. Linkedin users usually formally list all their occupation experiences in the "Experience" block, as illustrated in Figure 2. According to our statistics, on average, each user in our dataset has 5.65 occupations. For each user, we extracted his/her occupation list to serve as the ground truth. In this way, we obtained 12, 409 unique occupation titles. To avoid the typos and unusual occupation titles, such as "meeting booster," we filtered out occupation titles that occur fewer than 5 times, and we have 385 left. It is noteworthy that the vocabulary gap is very large for the occupation representation. Users with diverse backgrounds utilize various phrases to represent their titles. For instance, "programmer" and "software developer" are employed by different users to refer to the same "software engineer" title. We manually merged some variants. Ultimately, we have 80 occupation titles. Table II displays 20 representative occupations. Correspondingly, Figure 3 illustrates the user frequency distribution over these 20 occupations.

4. USER ATTRIBUTES LEARNING MODEL

To mathematically formulate our problem, we first define some notations. In particular, we use bold capital letters (e.g., $\mathbf{X}$) and bold lowercase letters (e.g., $\mathbf{x}$) to denote matrices and vectors, respectively. We employ lightface letters (e.g., x) to represent scalars and Greek letters (e.g., λ) as parameters. Unless stated otherwise, all vectors are in column form.

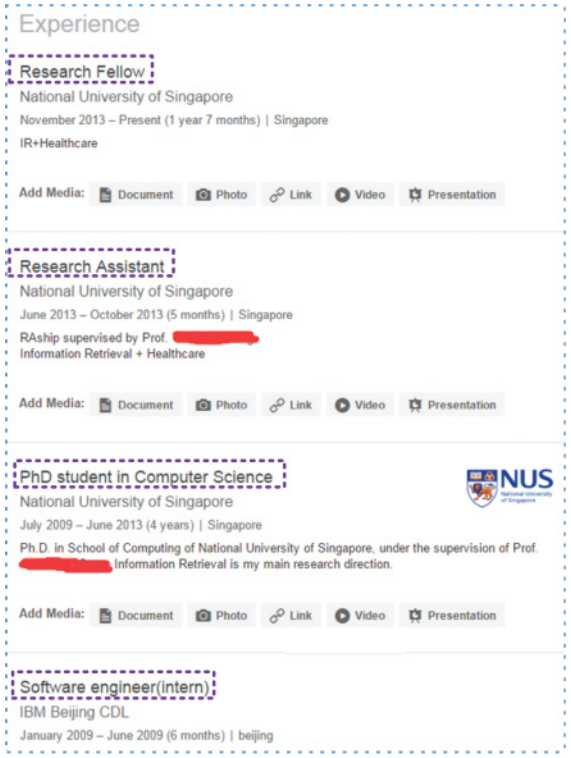

Fig. 2. Demonstration of working experiences listed in Linkedin, that is, occupation titles. They are highlighted by purple dotted rectangles. Due to the privacy concerns, we concealed some personal information.

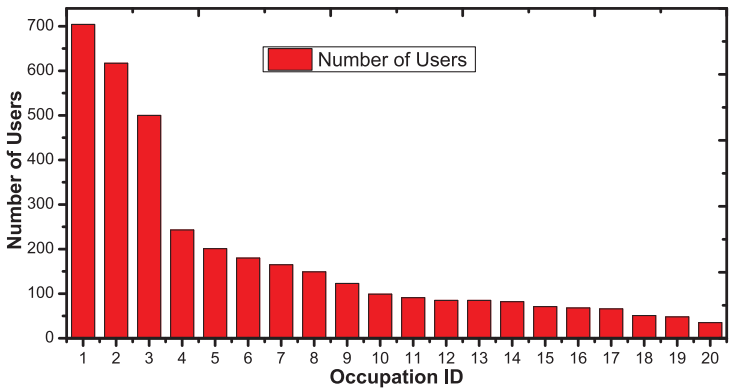

Fig. 3. Illustration of occupation-user distributions in our dataset. The occupation ID is aligned with the occupation title in Table II. We have 80 occupations, but we do not display them all due to space restrictions.

Assume that we are given N users $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T$ in the training set and their corresponding occupation categories $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_C] \in \mathbb{R}^{N \times C}$. Each user is characterized by V complementary sources from V mobile media platforms. For example, the nth user can be represented by $\mathbf{x}_n = [\mathbf{x}_{n1}^T, \mathbf{x}_{n2}^T, \ldots, \mathbf{x}_{nV}^T]^T$, where $\mathbf{x}_{nv} \in \mathbb{R}^{D_v}$, and D_v denotes the dimensionality of feature space on the vth source. All the training samples on the vth source and on all the sources are respectively represented

as $\mathbf{X}_v = [\mathbf{x}_{1v}, \mathbf{x}_{2v}, \ldots, \mathbf{x}_{Nv}]^T \in \mathbb{R}^{N\times D_v}$ and $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_v] \in \mathbb{R}^{N\times D}$, where $D = \sum_{v=1}^{V} D_v$. Our objective is to generalize the *gMM* model from training users to predict the occupations of other unseen users, given their information from multiple social media platforms.

We define $f_v^c(\mathbf{x}_{nv})$ as the probability function of user n falling into the cth occupation category, estimated based on the vth source. In a vectorwise form, the inference function of all users for the cth occupation by leveraging the vth source can be stated as

$$\mathbf{f}_v^c(\mathbf{X}_v) = \mathbf{X}_v\mathbf{w}_{cv}, \tag{1}$$

where $\mathbf{w}_{cv} \in \mathbb{R}^{D_v}$ is the coefficient vector for source v and task c. The probability for all users associated with cth occupation is modeled by averaging the prediction results from all sources,

$$\mathbf{f}^c(\mathbf{X}) = \frac{1}{V}\sum_{v=1}^{V}\mathbf{f}_v^c(\mathbf{X}_v) = \frac{1}{V}\sum_{v=1}^{V}\mathbf{X}_v\mathbf{w}_{cv}. \tag{2}$$

Based on the above definition, the traditional squared loss function that measures the empirical error on the training users can be formulated as $\sum_{c=1}^{C} \parallel \mathbf{y}_c - \frac{1}{V}\sum_{v=1}^{V}\mathbf{X}_v\mathbf{w}_{cv} \parallel^2$. As reported in Xu et al. [2012], the squared loss usually yields good performance as the other complex loss functions. We thus adopt the squared loss as the loss function in our algorithm for simplicity and efficiency.

To boost the occupation inference performance, besides the loss function, we simultaneously consider these three assumptions:

—***Source Consistency.*** We assume that the heterogeneous social media platforms of the same users characterize their behaviors from multiple views but should consistently reflect the same occupation type. Mathematically, $\mathbf{f}_v^c(\mathbf{X}_v)$ should be the same or very close with $\mathbf{f}_u^c(\mathbf{X}_u)$, for $u \neq v$.

—***Graph-Regularized Relatedness.*** It is reasonable to assume that some of the occupations are often more closely related and more likely to share common relevant input features than other occupations. We assume that occupations are related in a complex manner in the form of a graph. It is noteworthy that tree and some other basic structures are the special cases of graph structure. The graph structure can be built based on the external and internal prior knowledge.

—***Feature Selection.*** Feature sparsity is another reasonable assumption since only a small fraction of features are associated with their corresponding occupations. Feature selection using the ℓ_1 penalty (also referred to as 1-norm or Lasso penalty) has been shown to perform well when there are spurious features mixed with relevant features, which enforces many parameters being zeros and the parameter vector is thus sparse. Lasso and its advanced variants, such as fused lasso, should be considered to select discriminative features.

By jointly considering these assumptions, the objective function $\Phi(\mathbf{w}_{cv})$ of occupation inference can be formulated as

$$\begin{aligned}\min_{\mathbf{w}_{cv}} \frac{1}{2}\sum_{c=1}^{C}\left\| \mathbf{y}_c - \frac{1}{V}\sum_{v=1}^{V}\mathbf{X}_v\mathbf{w}_{cv}\right\|^2 + \frac{\lambda}{2}\sum_{c=1}^{C}\sum_{v=1}^{V}\sum_{v'\neq v}^{V} \parallel \mathbf{X}_v\mathbf{w}_{cv} - \mathbf{X}_{v'}\mathbf{w}_{cv'} \parallel^2 \\ + \beta \sum_{e=(c,c')\in\mathcal{E}} r_{cc'}\sum_{v=1}^{V} \parallel \mathbf{w}_{cv} - \mathbf{w}_{c'v} \parallel_1 + \mu\sum_{c=1}^{C}\sum_{v=1}^{V} \parallel \mathbf{w}_{cv} \parallel_1 .\end{aligned} \tag{3}$$

$$\min_{\mathbf{w}_{st}} \frac{1}{2}\sum_{t=1}^{T}\left\|\mathbf{y}_t - \frac{1}{S}\sum_{s=1}^{S}\mathbf{X}_s\mathbf{w}_{st}\right\|^2 + \frac{\lambda}{2}\sum_{t=1}^{T}\sum_{s=1}^{S}\sum_{s'\neq s}^{S} \parallel \mathbf{X}_s\mathbf{w}_{st} - \mathbf{X}_{s'}\mathbf{w}_{s't} \parallel^2$$
$$+ \beta \sum_{<t,t'>\in\mathcal{E}} r_{tt'}\sum_{s=1}^{S} \parallel \mathbf{w}_{st} - \mathbf{w}_{st'} \parallel_1 + \mu\sum_{t=1}^{T}\sum_{s=1}^{S} \parallel \mathbf{w}_{st} \parallel_1 . \quad (4)$$

The first term is the widely adopted least-squares loss function; the second term controls the source consistency; the third one implements the graph-guided fused lasso for multi-task classification that exploits the graph structure over the output variables; while the last term controls the sparsity. λ and β are parameters that respectively regularize the disagreement of heterogeneous sources for the same task and differences between related tasks on the same source. μ is a parameter that regulates the strength of the ℓ_1-norm regularization on the multi-source multi-task learning function.

Graph construction is the key to our proposed graph-guided fused lasso penalty. In the desired graph, each node represents one task or occupation, and each edge connects two nodes with its weight indicating the strength of correlation between two nodes. Two tasks connected by an edge with a high weight tend to be influenced by the same set of features. To construct the graph, we leveraged two kinds of prior knowledge encoded in external and internal resources. For the external resource, we exploited the web information indexed by Google. In particular, we treated each occupation title as a query and submitted it to the Google Search Engine. We collected the top 20 webpages for each occupation and then utilized the BoilerPipe tool[11] to extract the clean content from the pages. The similarities between pairwise occupations was estimated by computing the corresponding pairwise document similarities. In addition to the external resource, we also explored the internal knowledge embedded in our collected dataset. To be more specific, we measured the co-occurrence in users' Linkedin profiles in our dataset. For the given occupation c and c', let us denote $\mathcal{U}_c$ and $\mathcal{U}_{c'}$ as the sets of Linkedin profiles that respectively contain occupations c and c'. Inspired by the Jaccard coefficient [Lee et al. 2013], the relationship between these two occupations can be estimated by $\frac{|\mathcal{U}_c \cap \mathcal{U}_{c'}|}{|\mathcal{U}_c \cup \mathcal{U}_{c'}|}$. For pairwise occupations, their ultimate relationship was calculated by linearly averaging the similarity scores on the external and internal resources. We utilized $\mathbf{R}$ to denote the adjacency matrix of occupations, with its entry $r_{cc'}$ representing the relatedness between occupation c and c'. To avoid the commonly seen occupations overwhelming the non-commonly seen ones, we symetrically normalized $\mathbf{R}$ as $\mathbf{D}^{-\frac{1}{2}}\mathbf{R}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ is a diagonal matrix whose (i,i)-th element is the sum of the ith row of $\mathbf{R}$.

5. OPTIMIZATION

Our objective function $\Phi(\mathbf{w}_{cv})$ is convex, but the last two terms are non-smooth. Let us denote the last two terms as $\Psi_0(\mathbf{w}_{cv})$. We have

$$\Psi_0(\mathbf{w}_{cv}) = \beta \sum_{e=(c,c')\in\mathcal{E}} r_{cc'}\sum_{v=1}^{V} \parallel \mathbf{w}_{cv} - \mathbf{w}_{c'v} \parallel_1 + \mu\sum_{c=1}^{C}\sum_{v=1}^{V} \parallel \mathbf{w}_{cv} \parallel_1$$
$$= \beta \parallel \mathbf{WH} \parallel_1 + \mu \parallel \mathbf{W} \parallel_1 = \parallel \mathbf{WB} \parallel_1, \quad (5)$$

where $\mathbf{B} = [\beta\mathbf{H}, \mu\mathbf{I}] \in \mathbb{R}^{C\times(C+|\mathcal{E}|)}$ and $\mathcal{E}$ is the graph edge set; $\mathbf{W}$ is a block matrix, $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_C] \in \mathbb{R}^{D\times C}$, and $\mathbf{w}_c = [\mathbf{w}_{c1}^T, \mathbf{w}_{c2}^T, \ldots, \mathbf{w}_{cV}^T]^T$. $\mathbf{H} \in \mathbb{R}^{C\times|\mathcal{E}|}$ is a variant

[11] https://code.google.com/p/boilerpipe/.

vertex-edge incident matrix defined as follows:

$$H_{i,e} = \begin{cases} r_{cc'} & \text{if } e = (c, c') \text{ and } i = c; \\ -r_{cc'} & \text{if } e = (c, c') \text{ and } i = c'; \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Since the dual norm of the entrywise matrix ℓ_∞ norm is the ℓ_1 norm,[12] we can further restate the overall penalty in Equation (5) as:

$$\Psi_0(\mathbf{w}_{cv}) = \| \mathbf{WB} \|_1 \equiv \max_{\|\mathbf{A}\|_\infty \leq 1} tr(\mathbf{A}^T \mathbf{WB}), \tag{7}$$

where $\| \cdot \|_\infty$ is the matrix entrywise ℓ_∞ norm, defined as the maximum absolute value of all entries in the matrix, $\mathbf{A} \in \mathcal{P} = \{\mathbf{A} | \|\mathbf{A}\|_\infty \leq 1, \mathbf{A} \in \mathbb{R}^{D \times (C + |\mathcal{E}|)}\}$ is an auxiliary matrix associated with $\| \mathbf{WB} \|_1$, and $tr(\cdot)$ is the trace of a matrix.

The formulation in Equation (7) is still a non-smooth function with respect to $\mathbf{w}_{cv}$, which leads to a tough challenge for the optimization. Towards this end, we construct an approximation of Equation (7) with manageable errors. In particular, we define $\Psi_\gamma(\mathbf{w}_{cv})$ as

$$\Psi_\gamma(\mathbf{w}_{cv}) = \Psi_0(\mathbf{w}_{cv}) + \frac{\gamma}{2} \| \mathbf{A} \|_F^2 . \tag{8}$$

Let ϵ denote the approximation error. We then have

$$\epsilon = \Psi_\gamma(\mathbf{w}_{cv}) - \Psi_0(\mathbf{w}_{cv}) = \frac{\gamma}{2} \| \mathbf{A} \|_F^2 \tag{9}$$

$$\leq \max_{\|\mathbf{A}\|_\infty \leq 1} \frac{\gamma}{2} \| \mathbf{A} \|_F^2 \tag{10}$$

$$= \frac{\gamma}{2} D \times (C + |\mathcal{E}|). \tag{11}$$

We hence set the parameter γ as $\frac{2\epsilon}{D \times (C + |\mathcal{E}|)}$ to control the errors and to achieve the best convergence rate. In this work, we utilize $\Psi_\gamma(\mathbf{w}_{cv})$ to approximate $\Psi_0(\mathbf{w}_{cv})$.

By taking the derivative of $\Psi_\gamma(\mathbf{w}_{cv})$ over $\mathbf{A}$ and setting it to zero, we can derive $\mathbf{A} = \frac{\mathbf{WB}}{\gamma}$. We then apply $\mathbf{A}$ to $\mathcal{P}$, and we reach the optimal solution of $\mathbf{A}$,

$$\mathbf{A}^* = \Gamma\left(\frac{\mathbf{WB}}{\gamma}\right), \tag{12}$$

where Γ is a mapping function. It projects any $x \in \mathbb{R}$ into

$$\Gamma(x) = \begin{cases} 1 & \text{if } x > 1, \\ x & \text{if } -1 \leq x \leq 1, \\ -1 & \text{if } x < -1. \end{cases}$$

For a given matrix $\mathbf{A}$, $\Gamma(\mathbf{A})$ is defined as applying Γ on each of the entries in $\mathbf{A}$.

$\Psi_\gamma(\mathbf{w}_{cv})$ is a convex and continuously differentiable function with respect to $\mathbf{w}_{cv}$ for any $\gamma > 0$. We treat $\Psi_\gamma(\mathbf{w}_{cv})$ as a smooth approximation of $\Psi_0(\mathbf{w}_{cv})$. In particular, we have

$$\frac{\partial \Psi_\gamma(\mathbf{w}_{cv})}{\partial \mathbf{w}_{cv}} = \frac{\partial tr(\mathbf{A}^{*T} \mathbf{WB})}{\partial \mathbf{w}_{cv}} = \frac{\partial tr(\mathbf{WBA}^{*T})}{\partial \mathbf{w}_{cv}}. \tag{13}$$

[12] https://wiki.sfu.ca/personal/aoberman/index.php/Norms_and_dual_norms.

Let us denote $\mathbf{Q} = \mathbf{B}\mathbf{A}^{*T}$, where $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_C]^T \in \mathbb{R}^{C\times D}$ and $\mathbf{q}_c = [\mathbf{q}_{c1}^T, \mathbf{q}_{c2}^T, \ldots, \mathbf{q}_{cV}^T]^T$. We then can restate the above formulations as

$$\begin{aligned}\frac{\partial tr(\mathbf{W}\mathbf{Q}^T)}{\partial \mathbf{w}_{cv}} &= \frac{\partial tr\left(\sum_{c=1}^{C}\mathbf{w}_c\mathbf{q}_c^T\right)}{\partial \mathbf{w}_{cv}} = \frac{\partial tr(\mathbf{w}_c\mathbf{q}_c^T)}{\partial \mathbf{w}_{cv}} \\ &= \frac{\partial tr(\mathbf{q}_c^T\mathbf{w}_c)}{\partial \mathbf{w}_{cv}} = \frac{\partial tr\left(\sum_{v=1}^{V}\mathbf{q}_{cv}^T\mathbf{w}_{cv}\right)}{\partial \mathbf{w}_{cv}} \\ &= \mathbf{q}_{cv}^T.\end{aligned} \tag{14}$$

Taking the derivative of our objective function $\Phi(\mathbf{w}_{cv})$ in Equation (4) over $\mathbf{w}_{cv}$, we obtain

$$\frac{\partial \Phi(\mathbf{w}_{cv})}{\partial \mathbf{w}_{cv}} = \frac{1}{V}\mathbf{X}_v^T\left(\frac{1}{V}\sum_{v=1}^{V}\mathbf{X}_v\mathbf{w}_{cv} - \mathbf{y}_c\right) + \lambda\mathbf{X}_v^T\sum_{v'\neq v}^{V}(\mathbf{X}_v\mathbf{w}_{cv} - \mathbf{X}_{v'}\mathbf{w}_{cv'}) + \mathbf{q}_{cv}^T. \tag{15}$$

By rearranging the term orders of Equation (15), we arrive at

$$\frac{1}{V}\mathbf{X}_v^T\mathbf{y}_c - \mathbf{q}_{cv}^T = \left\{\frac{1}{V^2}\mathbf{X}_v^T\mathbf{X}_v + \lambda(V-1)\mathbf{X}_v^T\mathbf{X}_v\right\}\mathbf{w}_{cv} + \left(\frac{1}{V^2} - \lambda\right)\mathbf{X}_v^T\sum_{v'\neq v}^{V}\mathbf{X}_{v'}\mathbf{w}_{cv'}. \tag{16}$$

To facilitate the optimization analysis, we define some notations and rewrite Equation (16) in the following form:

$$\mathbf{L}\mathbf{w}_c = \mathbf{t}^c. \tag{17}$$

The above formulation is equivalent to the following linear system:

$$\begin{bmatrix} \mathbf{L_{11}} & \mathbf{L}_{12} & \mathbf{L}_{13} & \cdots & \mathbf{L}_{1V} \\ \mathbf{L}_{21} & \mathbf{L}_{22} & \mathbf{L}_{23} & \cdots & \mathbf{L}_{2V} \\ \mathbf{L}_{31} & \mathbf{L}_{32} & \mathbf{L}_{33} & \cdots & \mathbf{L}_{3V} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{L}_{V1} & \mathbf{L}_{V2} & \mathbf{L}_{V3} & \cdots & \mathbf{L}_{VV} \end{bmatrix}\begin{bmatrix} \mathbf{w}_{c1} \\ \mathbf{w}_{c2} \\ \mathbf{w}_{c3} \\ \vdots \\ \mathbf{w}_{cV} \end{bmatrix} = \begin{bmatrix} \mathbf{t}_1^c \\ \mathbf{t}_2^c \\ \mathbf{t}_3^c \\ \vdots \\ \mathbf{t}_V^c \end{bmatrix}, \tag{18}$$

where each involved element is defined as

$$\begin{cases} \mathbf{t}_v^c &= \frac{1}{V}\mathbf{X}_v^T\mathbf{y}_c - \mathbf{q}_{cv}^T, \\ \mathbf{L}_{vv} &= \frac{1}{V^2}\mathbf{X}_v^T\mathbf{X}_v + \lambda(V-1)\mathbf{X}_v^T\mathbf{X}_v, \\ \mathbf{L}_{vv'} &= (\frac{1}{V^2} - \lambda)\mathbf{X}_v^T\mathbf{X}_{v'}. \end{cases} \tag{19}$$

$\mathbf{L}$ is a positive definite matrix and hence it is invertible [Song et al. 2015a, 2015b; Nie et al. 2015, 2016]. We can hence derive the closed-form solution of $\mathbf{w}_c$,

$$\mathbf{w}_c = \mathbf{L}^{-1}\mathbf{t}^c. \tag{20}$$

We iteratively calculate $\mathbf{W}$ and $\mathbf{A}$ until convergence. The initial values of $\mathbf{W}$ is generated by randomization. In the whole process of our iteration, each step decreases the objective function value $\Phi(\mathbf{w}_{cv})$, whose lower bound is zero and therefore the convergence of our model is guaranteed [Ma et al. 2011].

6. EXPERIMENTS

In this section, we conducted extensive evaluations to thoroughly verify our proposed model and each of its components.

6.1. Experimental Settings

6.1.1. Data Preprocessing. We observed that the textual part of our dataset is of varying quality. Some users may write meaningless tips at their check-in venues, and some suspicious users may leave spams in Twitter. Before feature extraction, we performed the following pre-processing steps on the textual parts to filter the noise and reduce the feature space:

(1) We converted all uppercase letters to lowercase ones, and all slang words to synonyms based on the external dictionary to reduce the number of possible terms;
(2) We eliminated texts with phone numbers or email addresses since they are generally spams;
(3) We removed all the non-alphanumeric characters to filter out the meaningless numbers;
(4) We removed stop words and those terms with frequencies less than five.

6.1.2. Evaluation Metrics. For the task of user occupation inference, precision is of more importance as compared to recall. We thus validated our model via two widely accepted metrics that are able to capture precisions from different aspects. The first one is average S@K over all testing users. S@K measures the probability of finding a correct occupation title among the top K predicted ones. To be more specific, for each testing user, S@K is assigned to be 1 if a correct occupation is ranked in the top K positions and 0 otherwise. The second one is average P@K. P@K stands for the proportion of recommended occupations that are true. P@K is defined as

$$P@K = \frac{|\mathcal{C} \cap \mathcal{T}|}{|\mathcal{C}|}, \tag{21}$$

where $\mathcal{C}$ is a set of the top K predicted occupations and $\mathcal{T}$ is the set of true ones based on the ground truth.

6.1.3. Data Partition for Training and Testing. Considering the small size of our constructed dataset, the experimental results reported in this article were based on 10-fold cross-validation. In particular, we broke data into 10 sets of size 318 users per set. We trained our model on nine sets and tested it on one set each time. We repeated this process 10 times and took a mean accuracy as the ultimate result.

6.2. Feature Extraction

6.2.1. Textual Features for Twitter and Foursquare. After textual data preprocessing, we extracted and normalized the following features:

—**User Topics.** According to our observation, users may have higher probabilities of talking about topics related to their occupations. This motivates us to explore the topic distributions of users' social posts to characterize their occupations. For each user, we merged his/her tweets (Foursquare tips) into one document to represent the Twitter (Foursquare) source. We generated topic distributions using Latent Dirichlet Allocation [Blei et al. 2003], which has been widely found to be useful in latent topic modeling [Ji et al. 2015; Zhao et al. 2014]. Based on the metric of perplexity [Li et al. 2010] that is frequently utilized to find the optimal number of hidden topics, we ultimately obtained 50- and 30-dimensional (50D and 30D) topic-level features over users' tweets in Twitter and tips in Foursquare, respectively.
—**Linguistic Inquiry and Word Count features (LIWC).** LIWC is widely used to analyze the psycho-linguistic transparent lexicon. It plays an important role in predicting users' personality and careers [Lee et al. 2015]. The main component of

LIWC is a directory that contains the mapping from words to 72 categories.[13] Given a document, LIWC computes the percentage of words in each category and represents it as a vector of 72 dimensions.

—**Heuristically Inferred Features.** Inspired by Farseev et al. [2015], we extracted some heuristically inferred features. First, we counted the number of URLs, number of hash tags, and number of user mentions, since these features are correlated with users' social network activity level and can thus indicate users' occupation types. Second, we counted the number of slang words, number of emotion words,[14] and number of emoticons and computed an average sentiment score. These features can be good signals of user personality traits, which in turn are occupation dependent. Third, we computed some writing behavior style features, including "number of repeated characters" in words, "number of misspellings," and "number of unknown to the spell checker words," which are often reflected by users' occupations. In this way, we extracted 14D features.

6.2.2. Semantic Location Features for Foursquare. Besides the topic features extracted from the tips, we also extracted the semantic location features for the source of Foursquare. In particular, we utilized two attributes related to the check-in behaviors. First, we collected the well-structured and hierarchically organized venue categories of Foursquare.[15] Each Foursquare venue is mapped to one or more categories depending on its social function. Second, users visit venues at different times, which shows the temporal dimensions related to users' behaviors. To explore the semantics of spatial and temporal information, we represent each user by a weighted vector, where each dimension represents a visit to a particular venue category at a particular time period. In total, we utilized 423 leaf categories and the eight different time periods, which are {morning (5am–11am), afternoon (12pm–18pm), evening (19pm–23pm), night (12am–4am)} $\times$ {weekday, weekend}.

6.2.3. Visual Features for Instagram. To represent the content of each image, we extracted the following features:

—**Local Features.** We used the difference of Gaussians to detect keypoints in each image and extracted their SIFT (Scale-Invariant Feature Transform) descriptors. By building a visual codebook of size 1,000 based on k-means, we obtained a 1,000D bag-of-visual-words histogram for each image [Ji et al. 2013].

—**Global Features.** We further extracted 428D global visual features, including 225D blockwise color moments based on 5×5 fixed partition of the image, 128D wavelet texture, and 75D edge direction histogram [Nie et al. 2012].

In summary, we extracted 136D, 547D, and 1,428D features for each given user from their Twitter, Foursquare, and Instagram sources, respectively. We fed the features into our model for validation.[16]

6.3. Overall Model Evaluation

To demonstrate the effectiveness of our proposed user attributes learning model, we comparatively verified the following state-of-the-arts competitors:

[13] http://www.liwc.net/.

[14] www.sentiwordnet.isti.cnr.it.

[15] https://developer.foursquare.com/categorytree.

[16] Actually, we also investigated some other possible features, such as some n-grams and part-of-speech tags, but these features do not have strong relation with occupation inference, and hence a detailed description and evaluation of such features are omitted in this article.

Table III. Performance Comparison among Various Models for Occupation Inference from Multiple Social Media Platforms. The Performance Is Measured in Terms of *S@K* and *P@K* (%)

Models	S@1	S@2	S@3	S@4	S@5	P@1	P@2	P@3	P@4	P@5
SVM	41.04	56.26	66.29	72.67	75.38	41.04	28.13	22.10	18.17	15.08
GLR	46.73	60.53	69.81	75.97	77.26	46.73	30.27	23.27	18.99	15.45
MSIF	50.31	63.46	72.33	79.18	82.30	50.31	31.73	24.11	19.80	16.46
MTL	51.29	65.09	73.58	78.58	81.76	51.29	32.55	24.53	19.65	16.35
regMVMT	55.97	68.84	77.92	85.19	88.02	55.97	34.42	25.97	21.30	17.60
gMM	**60.50**	**72.58**	**82.04**	**88.46**	**92.77**	**60.50**	**36.29**	**27.35**	**22.11**	**18.55**

(1) *SVM*: SVM is a typical mono-source mono-task learning model [Chang and Lin 2011]. We concatenated all the features from multiple sources into a single vector and trained each task separately. With the assist of LIBSVM (A Library for Support Vector Machines),[17] we chose radial-basis as our kernel function.
(2) *GLR*: The Group Lasso Regularization method [Chapelle et al. 2014; Meier et al. 2008] is an $\ell_{2,1}$-norm penalty for group feature selection $\frac{1}{2}\|\mathbf{XW}-\mathbf{Y}\|_F^2+\rho\|\mathbf{W}\|_F^2+\sigma\|\mathbf{W}\|_{2,1}$. This model encodes the group sparsity but fails to take the task relatedness and source relatedness into account.
(3) *MSIF*: The authors in Huang et al. [2014] proposed a Multi-Source Integration Framework to infer users' occupation from their social activities recorded in the social sites, which combines both content model and network model. As reported in Huang et al. [2014], this work outperforms most of the prevailing occupation inference approaches. That is why we did not selected others as competitors.
(4) *MTL*: As a typical example of the traditional multi-task learning model, the work in Zhang and Yeung [2010] aims to automatically capture and model the task relatedness. It is formulated as $\frac{1}{2}\|\mathbf{XW}-\mathbf{Y}\|_F^2+\upsilon\|\mathbf{W}\Omega^{-1}\mathbf{W}^T\|_F^2+\varrho\|\mathbf{W}\|_F^2$. The code for this model is available here.[18]
(5) *regMVMT*: A semi-supervised inductive multi-view multi-task learning model introduced in Zhang and Huan [2012]. As reported in Zhang and Huan [2012], this model regulates both the source consistency and the task relatedness. However, it simply assumes and characterizes the uniform relatedness among tasks.
(6) *gMM*: Our proposed graph-guided multi-task multi-source learning model.

We can see that the selected competitors are very comprehensive, including mono-source mono-task learning, group lasso regularization, multi-task learning, multi-source learning, as well as multi-source multi-task learning. For each method mentioned above, the involved parameters were carefully tuned with 5-fold cross validation in the training data between 10^{-2} and 10^2, and the parameters with the best performance with respect to P@5 were used to report the final results.

The comparison results of various models for occupation inference from multiple social media sources are shown in Table III. Experimental results are measured by *P@K* and *S@K*, respectively. From this table, we have some observations: (1) *SVM* achieves the worst performance. One possible reason may be the insufficient positive training samples for certain occupation titles. For instance, only 18 positive training samples are available for the occupation "musician." (2) *GLR* is slightly better than *SVM* but still worse than others. In this model, the weights of each feature over all tasks are grouped using the ℓ_2 norm, and all features are further grouped using the ℓ_1 norm. This model thus tends to select features based on the strength of the feature over all tasks. Its result confirms that some features have little description power for

[17] http://www.csie.ntu.edu.tw/∼cjlin/libsvm.
[18] http://www.comp.hkbu.edu.hk/∼yuzhang/codes/MTRL.zip.

Table IV. The P-Value of Pairwise Significance Test between Our Proposed *gMM* Model and Each of the Baselines. We Can See That All the P-Values Are Significantly Smaller Than 0.05, and They Indicate That the Differences Are Significant

Metrics / Pairwise Significance Test	p-value over S@5	p-value over P@5
gMM VS. SVM	$2.82E$-05	$1.79E$-03
gMM VS. GLR	$5.32E$-04	$1.03E$-03
gMM VS. $MSIF$	$6.80E$-03	$5.31E$-02
gMM VS. MTL	$1.28E$-03	$4.08E$-02
gMM VS. $regMVMT$	$1.05E$-03	$2.07E$-02

all the tasks and hence feature selection is necessary. (3) *MSIF* and *MTL* show their superiors to *SVM* and *GLR*. The reason may be that they either explicitly consider the source fusion or consider the task relatedness. Appropriate source fusion can enhance user characterization, and multiple task learning can effectively increase the number of samples by learning multiple related tasks simultaneously. (4) *regMVMT* outperforms all the competitors, except *gMM*. This is because it jointly regularizes the task relatedness and source fusion. However, it restricts the relatedness among tasks in a uniform way without considering prior knowledge. Moreover, it is unable to select discriminative task-specific and task-sharing features. As an extension of *regMVMT*, our proposed *gMM* has well resolved these issues by leveraging the graph-guided fused lasso penalty. That is why it achieves the best performance consistently in terms of S@K and P@K at different depths. And (5) the performance in terms of S@1 and P@1 is as high as 60.50%, which means that for up to 60% of users, our model can precisely infer their occupations if our model only predicted one occupation for each user only. Meanwhile, the value of S@5 almost ensures that at leat one occupation is correct among the top five predicted ones.

We also conducted the analysis of variance based on S@5 and P@5, respectively. To be more specific, we performed a pairwise t-test between our proposed gMM model and each of the competitors based on 10-fold cross-validation results in terms of S@5 and P@5. The results are summarized in Table IV. It can be observed that all the p-values are much smaller than 0.05, which shows that the improvements of our proposed model over other baselines are statistically significant.

6.4. Component-wise Analysis

To examine how effective each component is in the proposed gMM model, we carried out experiments to compare the performance of the following methods. In fact, these methods can be deduced from our gMM model by excluding some terms:

(1) *SLF*: The Squared Loss Function can be derived by setting λ, β, and μ to be zero, respectively. In this case, we do not consider both the source relatedness and task relatedness.
(2) *MSL*: The Multi-Source Learning model can be obtained by setting β as zero. In such a case, the graph-guided fused lasso is not considered.
(3) *MTLg*: The Multi-Task Learning model with graph-guided penalty is a special case of our proposed gMM model by setting λ to be zero [Chen et al. 2010]. In such a context, the source relatedness is not taken into consideration.
(4) gMM: Our proposed multi-source learning model with the graph-guided fused lasso penalty.

The results of component-wise analysis are summarized in Table V. From this table, the following observations can be made: (1) *SLF* is the most unsatisfactory method,

Table V. Experimental Results of Component-Wise Analysis. The Experimental Results Are Measured by $S@K$ and $P@K$, Respectively (%)

Components	S@1	S@2	S@3	S@4	S@5	P@1	P@2	P@3	P@4	P@5
SLF	39.53	56.79	63.27	68.05	71.04	39.53	28.40	21.09	17.01	14.21
MSL	52.67	64.06	73.84	80.72	84.91	52.67	32.03	24.61	20.18	16.98
MTLg	54.09	67.61	75.60	82.23	87.45	54.09	33.81	25.20	20.56	17.49
gMM	**60.50**	**72.58**	**82.04**	**88.46**	**92.77**	**60.50**	**36.29**	**27.35**	**22.11**	**18.55**

Table VI. Experimental Results of Source Combination Evaluation. The Experimental Results Are Measured by $S@K$ and $P@K$, Respectively (%). The Last Column Displays the Results of Pairwise Significance Test between the Model Leveraging All Three Sources and Those Only Leveraging Parts of Them (Based on 10-Fold Cross-Validation Results in Terms of P@5.)

Source Combination	S@1	S@3	S@5	P@1	P@3	P@5	p-value
Twitter	57.55	79.34	90.72	57.55	26.45	18.14	1.06*E*-02
Instagram	35.50	56.98	63.27	35.50	18.99	12.65	4.26*E*-06
Foursquare	41.86	62.80	70.60	41.86	20.93	14.12	3.41*E*-05
Twitter + *Instagram*	58.02	79.91	81.60	58.02	26.64	16.32	1.87*E*-02
Twitter + *Foursquare*	59.03	80.91	91.82	59.03	26.97	18.36	2.45*E*-02
Instagram + *Foursquare*	48.05	68.62	74.53	48.05	22.87	14.91	8.26*E*-04
Twitter + *Instagram* + *Foursquare*	**60.50**	**82.04**	**92.77**	**60.50**	**27.35**	**18.55**	−−

whose result is even worse than that of *SVM*, as shown in Table III. This may be caused by three facts. First, the model of *SLF* does not support the function of sparsity or complexity control, and hence it is easy to overfit. Second, it neither explores the structural relatedness among tasks, nor considers the agreement among sources. Third, it is unable to select descriptive features. The comparative results imply that feature selection, multi-source learning, and multi-task learning are of vital importance for user attributes learning. (2) The performance of *MSL* is much better than *SLF*, but it is less promising as compared to that of *MTLg*. This demonstrates that the graph-guided fused lasso penalty holds more encouraging effects than that of multi-source fusion. The main reason may be that our dataset is not very large, and *MTLg* enables the training samples sharing among closely related tasks, which greatly alleviates the problem of insufficient training samples. (3) Jointly analyzing Tables III and V, we can see that *MTLg* performs better than the conventional *MTL*. The possible reason is that, besides task relatedness modeling, the former also plays a role in feature selection. And (4) our proposed *gMM* seamlessly sews all these components and stably works best. We also conducted the the pairwise significance test between *gMM* and each of its components. All the p-values are smaller than 0.05, which shows that the improvements boosted by *gMM* are statistically significant.

6.5. Source Integration

We believe that the discriminative capabilities for individual source or source combinations vary significantly. We thus conducted experiments to study the representativeness of individual sources and various source combinations for user occupation inferences.

Table VI comparatively displays the experimental results. Notably, when learning on individual sources, our proposed *gMM* model degenerates into a multi-task learning model with the graph-guided penalty, which only considers the task relatedness and feature selection. It is intuitive that feeding our model with all three sources obtains the best results. Meanwhile, we noticed that the model trained on individual Twitter is effectively ahead of that trained on individual Foursquare and Instagram, respectively. This reveals that the Twitter source is much more informative in the user representation. One reason may be that the users in our dataset are very active in Twitter, and

the data crawled from Twitter are hence very intensive. We also observed that in the bi-source combinations, the integration of Twitter and Foursquare outperforms others but is still suboptimal as compared to the tri-source combination. This tells us that the visual information from Instagram is a positive added value. From the last column of Table VI, it can be seen that the model leveraging all the three sources is statistically better than those leveraging only parts of them.

6.6. Parameter Tuning

Our model holds three key parameters as shown in Equation (4). The optimal values of these parameters were carefully tuned with 5-fold cross-validation in the training data. In particular, based on the 10-fold cross-validation, we have around 2,862 users during each training round. We performed 5-fold cross-validation on the 2,862 users to learn the optimal parameters by grid search strategy between 10^{-2} and 10^{2} with small but adaptive step size. The step sizes were 0.01, 0.05, 1, and 5 for the range of [0.01, 0.1], [0.1, 1], [1,10], and [10,100], respectively. The parameters corresponding to the best p@5 were used to report the final results. For other competitors, the procedures to tune the parameters are analogous to ensure fair comparison.

6.7. Computational Analysis

In the training process, the computational complexity comes from three parts: (1) graph construction and normalization, $O(C^3)$; (2) calculation of $\mathbf{A}$ and $\mathbf{W}$, $O(D \times (C + |\mathcal{E}|))$ and $O(D^3 + CD^2)$, respectively; and (3) the alternative iteration between $\mathbf{A}$ and $\mathbf{W}$, p times. Therefore, C, D, $|\mathcal{E}|$, and p respectively refer to the number of occupation titles studied in our work, the feature dimensions extracted from all the sources, number of edges in the constructed graph, and iteration times. Hence, the overall time complexity scales as $O(C^3 + p(D^3 + CD^2 + CD + D|\mathcal{E}|))$. Usually, D is much larger than C, and the final complexity can be thus restated as $O(pD^3)$. In our work, C is 80, D is 2, 111, and p is in the order of hundreds. The process can be completed in less than 10s excluding feature extraction over a computer equipped with a 3.4GHZ CPU (Central Processing Unit) with 16GB RAM (Random Access Memory). Therefore, our model has a large potential to be applied to other web-scale or real-time applications.

7. CONCLUSION AND FUTURE WORK

This article has presented a novel model for user attributes learning from multiple mobile social media platforms. This model seamlessly sews complementary information from heterecious sources and regularizes the inter-task relatedness with the graph-guided penalty. In addition, it also jointly learns the task-sharing and task-specific features via lasso and graph-guided fused lasso. The graph-structure is constructed by leveraging the external and internal knowledge. We relaxed the non-smooth objective function to a smooth one and derived its optimization process. To validate our proposed model, we crawled ground truth, short texts, check-ins, and images from four prevailing sources, namely Linkedin, Twitter, Foursquare, and Instagram. Representative features were extracted from these sources to characterize users from various views. Extensive experiments on this dataset show the priority of our model to the baselines. It is worth mentioning that our model is flexible to incorporate more sources and it is applicable to infer other attributes beyond occupations.

In future, we will extend our model from two angles. First, in the current model, we do not consider the contribution confidences of various sources. We plan to adaptively learn the source weights. Second, we will rebuild our model to automatically learn the penalty structure.

REFERENCES

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* (2003).

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* (2011).

Olivier Chapelle, Eren Manavoglu, and Romer Rosales. 2014. Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology* (2014).

Xi Chen, Seyoung Kim, Qihang Lin, Jaime G. Carbonell, and Eric P. Xing. 2010. Graph-structured multi-task regression and an efficient optimization method for general fused lasso. *Comput. Res. Repos.* (2010).

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011. Political polarization on twitter. In *Proceedings of the International Conference on Weblogs and Social Media*.

Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, and Nitesh V. Chawla. 2014. Inferring user demographics and social strategies in mobile social networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Alex Farseev, Liqiang Nie, Mohammad Akbari, and Tat-Seng Chua. 2015. Harvesting multiple sources for user profile learning: A big data study. In *International Conference on Multimedia Retrieval*.

Liang Gou, Michelle X. Zhou, and Huahai Yang. 2014. KnowMe and ShareMe: Understanding automatically discovered personality traits from social media and user sharing preferences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Jingrui He and Rick Lawrence. 2011. A graph-based framework for multi-task multi-view learning. In *Proceedings of the International Conference on Machine Learning*.

Yanxiang Huang, Lele Yu, Xiang Wang, and Bin Cui. 2014. A multi-source integration framework for user occupation inference in social media systems. *World Wide Web* 18, 5 (2014), 1247–1267.

Rongrong Ji, Ling-Yu Duan, Jie Chen, Tiejun Huang, and Wen Gao. 2014. Mining compact bag-of-patterns for low bit rate mobile visual search. *IEEE Trans. Image Process.* (2014).

Rongrong Ji, Ling-Yu Duan, Jie Chen, Lexing Xie, Hongxun Yao, and Wen Gao. 2013. Learning to distribute vocabulary indexing for scalable visual search. *IEEE Trans. Multimedia* 15, 1 (2013), 153–166.

Rongrong Ji, Yue Gao, Wei Liu, Xing Xie, Qi Tian, and Xuelong Li. 2015. When location meets social multimedia: A survey on vision-based recognition and mining for geo-social multimedia analytics. *ACM Trans. Intell. Syst. Technol.* (2015).

Xin Jin, Fuzhen Zhuang, Shuhui Wang, Qing He, and Zhongzhi Shi. 2013. Shared structure learning for multiple tasks with multiple views. In *Machine Learning and Knowledge Discovery in Databases*.

Kyumin Lee, Jalal Mahmud, Jilin Chen, Michelle Zhou, and Jeffrey Nichols. 2015. Who will retweet this? Detecting strangers from twitter to retweet information. *ACM Trans. Intell. Syst. Technol.* (2015).

Yugyung Lee, Saranya Krishnamoorthy, and Deendayal Dinakarpandian. 2013. A semantic framework for intelligent matchmaking for clinical trial eligibility criteria. *ACM Trans. Intell. Syst. Technol.* (2013).

Daifeng Li, Bing He, Ying Ding, Jie Tang, Cassidy Sugimoto, Zheng Qin, Erjia Yan, Juanzi Li, and Tianxi Dong. 2010. Community-based topic modeling for social tagging. In *Proceedings of the ACM International Conference on Information and Knowledge Management*.

Zhigang Ma, Yi Yang, Feiping Nie, Jasper Uijlings, and Nicu Sebe. 2011. Exploiting the entire feature space with sparsity for automatic image annotation. In *Proceedings of the ACM International Conference on Multimedia*.

Lukas Meier, Sara van de Geer, and Peter Bühlmann. 2008. The group lasso for logistic regression. *J. Roy. Stat. Soc. Ser. B* 70, 1 (2008), 53–71.

Minh-Thap Nguyen and Ee-Peng Lim. 2014. On predicting religion labels in microblogging networks. In *Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval*.

Liqiang Nie, Xuemeng Song, and Tat-Seng Chua. 2016. *Learning from Multiple Social Networks*. Morgan & Claypool.

Liqiang Nie, Shuicheng Yan, Meng Wang, Richang Hong, and Tat-Seng Chua. 2012. Harvesting visual concepts for image search with complex queries. In *Proceedings of the ACM International Conference on Multimedia*.

Liqiang Nie, Luming Zhang, Yi Yang, Meng Wang, Richang Hong, and Tat-Seng Chua. 2015. Beyond doctors: Future health prediction from multimedia and multimodal observations. In *Proceedings of the Annual ACM Conference on Multimedia Conference*.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A machine learning approach to twitter user classification. In *Proceedings of the International Conference on Weblogs and Social Media*.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the International Workshop on Search and Mining User-generated Contents*.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Xuemeng Song, Liqiang Nie, Luming Zhang, Mohammad Akbari, and Tat-Seng Chua. 2015a. Multiple social network learning and its application in volunteerism tendency prediction. In *Proceedings of the ACM SIGIR Conference*.

Xuemeng Song, Liqiang Nie, Luming Zhang, Maofu Liu, and Tat-Seng Chua. 2015b. Interest inference via structure-constrained multi-source multi-task learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Qianqian Xu, Qingming Huang, and Yuan Yao. 2012. Online crowdsourcing subjective image quality assessment. In *Proceedings of the ACM International Conference on Multimedia*.

Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *Proceedigs of the International Conference on Weblogs and Social Media*.

Jintao Zhang and Jun Huan. 2012. Inductive multi-task learning with multiple view data. In *Proceedings of the International ACM SIGKDD Conference*.

Yu Zhang and Dit-Yan Yeung. 2010. A convex formulation for learning task relationships in multi-task learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Yi-Liang Zhao, Liqiang Nie, Xiangyu Wang, and Tat-Seng Chua. 2014. Personalized recommendations of locally interesting venues to tourists via cross-region community matching. *ACM Trans. Intell. Syst. Technol.* (2014).

Received June 2015; revised February 2016; accepted June 2016